



EMOE: Modality-Specific Enhanced Dynamic Emotion Experts

Yiyang Fang*, Wenke Huang*, Guancheng Wan*, Kehua Su†, Mang Ye†

National Engineering Research Center for Multimedia Software,
 School of Computer Science, Wuhan University.

{fangyiyang, wenkehuan, yemang}@whu.edu.cn

<https://github.com/fuyyyyy/EMOE>

Abstract

Multimodal Emotion Recognition (MER) aims to predict human emotions by leveraging multiple modalities, such as vision, acoustics, and language. However, due to the heterogeneity of these modalities, MER faces two key challenges: modality balance dilemma and modality specialization disappearance. Existing methods often overlook the varying importance of modalities across samples in tackling the modality balance dilemma. Moreover, mainstream decoupling methods, while preserving modality-specific information, often neglect the predictive capability of unimodal data. To address these, we propose a novel model, Modality-Specific Enhanced Dynamic Emotion Experts (EMOE), consisting of: (1) Mixture of Modality Experts for dynamically adjusting modality importance based on sample features, and (2) Unimodal Distillation to retain single-modality predictive ability within fused features. EMOE enables adaptive fusion by learning a unique modality weight distribution for each sample, enhancing multimodal predictions with single-modality predictions to balance invariant and specific features in emotion recognition. Experimental results on benchmark datasets show that EMOE achieves superior or comparable performance to state-of-the-art methods. Additionally, we extend EMOE to Multimodal Intent Recognition (MIR), further demonstrating its effectiveness and versatility.

1. Introduction

Multimodal Emotion Recognition (MER) is a critical task in affective computing that aims to predict human emotions by leveraging multiple data modalities, including vision, acoustics, and language. Compared to a single modality, different modalities provide unique and complementary in-

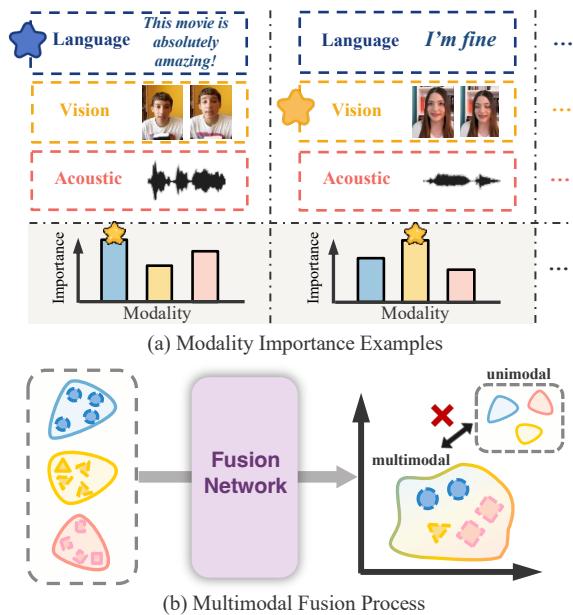


Figure 1. Illustration of modality importance examples and multimodal fusion process in MER. (a) illustrates a phenomenon where the modality content in different samples affects the **varying importance** of each modality during fusion. (In this paper, *sample* refers to the inputs from all modalities for a single instance.) (b) shows that during fusion, the model emphasizes multimodal invariant features while disregarding **unimodal ability**.

formation [54], which can enhance the prediction of emotional behaviors. With the development of deep learning [1, 19, 20, 59], MER has emerged as a rapidly expanding research field in affective computing [40], offering numerous applications, like intelligent tutoring systems [38], product feedback assessment [33], and robotics [31].

For MER, different modalities within the same video segment are often complementary, providing additional cues for semantic and emotional disambiguation. However, due to differences in how each modality captures and expresses information, as well as the varying amounts of

*Equal contributions

†Corresponding Author

information each modality contains across different video segments, there arises **modality balance dilemma** in multi-modal learning. Additionally, when different modalities are fused, there tends to be a focus on **invariant features** [30], which often leads to the neglect of modality-specific aspects, resulting in **modality specialization disappearance**.

Existing relative methods primarily focus on naively incorporating different modality information through popular combination techniques [36, 69], such as simple concatenation [14, 35], tensor fusion [63] and so on. Although previous work has attempted to address the modality balance dilemma using approaches such as modality imbalance regularization [11], multimodal contrastive learning[9], and hierarchical modality fusion [22, 53], these methods employ a unified fusion paradigm to process all samples, which fails to account for **① the varying importance of different modalities for distinct samples**. In fact, the differences between modalities in conveying information are more pronounced in emotion recognition compared to other multimodal learning tasks. As shown in Fig. 1(a), individuals express emotions inconsistently across different samples using all three modalities. For instance, one person may express emotions more through text and tone, while another might use facial expressions more predominantly. Consequently, different modalities possess varying abilities to predict emotions, and the significance of each modality during fusion fluctuates based on the specific input. On the other hand, to address modality specialization disappearance, related methodologies introduce various modality decoupling solutions to preserve modality-specific information [14, 29, 55], separating each modality into invariant and specific spaces before fusion. However, these methods often **② overlook the predictive capability of unimodal data**, as illustrated in Fig. 1(b). Although unimodal data may be less comprehensive, its stability and reliability can exceed that of multimodal data in certain contexts [16–18, 48]. This makes it a valuable addition to complex tasks, improving model robustness.

In this paper, we propose **Modality-Specific Enhanced Dynamic Emotion Experts (EMOE)**, which consists of two primary components. **First**, due to **①**, we consider designing a **method** that allows the model to autonomously learn the **modality importance distribution** for each sample. **Inspired** by the **Mixture of Experts** [3, 6], we introduce **Mixture of Modality Experts**. Here, a Router Network assesses the importance of each modality based on input features, treating each modality channel as an expert rather than a single network. Additionally, the dominant modalities tend to steer the learning process during model training, resulting in insufficient training of other modalities [36, 46, 47, 51] and an overestimation of the dominant modality’s weight. Therefore, we also propose an experts balancing mechanism based on router entropy loss to optimize the weight distribution and address modality dominance. **Second**, we

propose Unimodal Distillation with Router Selection to address **②**. In this approach, we incorporate a single-modality classification head into the model and calculate the prediction loss to ensure that the unimodal features possess predictive capabilities. Subsequently, we implement one-way distillation [10, 12, 15, 34] by aligning the multimodal predictive features with the unimodal features, weighted by their importance. This method allows the modality-specific features to guide the learning of multimodal features, ensuring that the prediction results consider both modality-invariant and modality-specific aspects.

We conducted relevant experiments to validate this approach. Given the similarities between Multimodal Intent Recognition (MIR) and MER [43, 57, 58], we applied our method to the MIR task. The main contributions of this work can be summarized as follows:

- We tackle the challenge of modality importance varying across samples in fusion. Using Mixture of Modality Experts, we derive a distinct weight distribution per sample, facilitating adaptive multimodal fusion.
- Recognizing that modality fusion often neglects single modality predictive ability, we propose Unimodal Distillation with Router Selection. This method uses the predictive information of each modality to guide the fused features, preserving modality-specific characteristics in the multimodal fusion.
- We conduct experiments on various datasets, including CMU-MOSI [62], CMU-MOSEI [65], and MIntRec [66], and achieve superior or comparable results to state-of-the-art methods. We also extend our approach from MER to MIR, validating the effectiveness of our method.

2. Related Works

2.1. Multimodal Emotion Recognition

Multimodal emotion recognition (MER) aims to leverage multiple data sources—such as visual, acoustic, and text modalities—to more accurately identify and classify human emotions. Previous work in MER [4, 56, 67, 69] can generally be categorized into two main directions: optimizing modality fusion and enhancing feature representation.

The first direction focuses on improving the integration of different modalities to ensure effective combination of complementary information [69]. Early fusion techniques primarily relied on simple operations like feature concatenation [35]. Zadeh *et al.* [63] later introduced tensor fusion, which enhanced performance by using trilinear tensor decomposition to capture high-order interaction features. More recently, the focus has shifted toward neural network-based methods, such as MuLT [44], which proposes a multimodal transformer incorporating a cross-modal attention mechanism to effectively learn the underlying adaptations and correlations between different modalities. However,

these methods typically apply the same paradigm to all input samples, neglecting the fact that the importance of each modality can vary depending on the specific sample. In a different line of research, Hazarika *et al.* [14] proposed decomposing multimodal features into modality-invariant and modality-specific components, enabling the learning of refined multimodal representations. Subsequent approaches addressing modality heterogeneity have primarily followed modality decoupling strategies. However, these methods often neglect individual modality strengths, while adding channels may cause conflicts and redundancies. Our method tackles this by dynamically weighting modalities and using single modalities to guide fused features, ensuring a balanced multimodal emotion prediction.

2.2. Imbalanced multimodal learning

Recent studies have highlighted the issue of imbalanced multimodal learning, where models tend to favor certain modalities over others, negatively impacting overall performance [21, 36]. Various strategies have been proposed to address this challenge, with the aim of balancing the optimization of individual modalities [8, 27, 49, 50]. Peng *et al.* [36] introduced a gradient modulation strategy that dynamically adjusts the contributions of different modalities during training, thereby reducing the influence of dominant modalities. However, in this approach, modalities with weaker expressive capabilities often hinder the learning process of stronger modalities. In response, Wei *et al.* [51] proposed "Diagnosing and Re-learning," adaptively re-training weaker modalities for better balance. In contrast, we tackle the issue from the sample level, recognizing imbalances across modalities for specific samples. When a dominant modality drives learning, it risks inaccurate predictions due to information bias. To solve this, we propose a dynamic modality weighting approach to optimize modality weights during training, preventing over-reliance on dominant modalities and preserving the predictive power of each modality, thus improving predictions.

2.3. Mixture-of-Experts

Mixture-of-Experts (MoE) replicates certain components of a network into multiple instances, known as experts [5, 23, 24]. Initially proposed to enhance model prediction performance, MoE leverages the cooperation and competition between multiple expert models. However, activating all experts increases computational costs. Sparse MoE (SMoE) [26, 28, 42, 70, 71] addresses this by activating only some experts, enhancing efficiency. SMoE is used in multi-task learning [6] for task-specific expert activation and in model compression [28] to cut computational demands. In multimodal learning, recent efforts replace dense layers with MoE structures to lower costs [68] or manage missing modalities [61]. Inspired by prior studies, we treat

networks for different modalities as experts and use a router network to learn their importance weights, enhancing data-level personalization in multimodal learning.

3. The Proposed Me

1. Preliminaries

Following the typical multimodal emotion recognition [14, 44], our goal is to detect sentiments in videos by leveraging multimodal signals. For a given utterance U , the input contains three sequences of low-level features corresponding to language (l), visual (v), and acoustic (a) modalities. These sequences are represented as $U_l \in \mathbb{R}^{T_l \times d_l}$, $U_v \in \mathbb{R}^{T_v \times d_v}$, $U_a \in \mathbb{R}^{T_a \times d_a}$. Here T_m represents the length of the utterance in modality m (e.g., T_l denotes the number of tokens in the language modality), and d_m corresponds to the dimensionality of the features for each modality. Given these sequences U_m , $m \in \{l, v, a\}$, the primary task is to predict the affective orientation of the utterance U . The prediction is either from a predefined set of C categories ($y \in \mathbb{R}^C$) or as a continuous intensity variable ($y \in \mathbb{R}$).

3.2. Mixture of Modality Experts

We consider three modalities: language (L), visual (V), and acoustic (A). Initially, three distinct 1D temporal convolutional layers are used to capture temporal patterns and extract raw features for each modality, denoted as $X_m \in \mathbb{R}^{T_m \times d_m}$, where $m \in \{l, v, a\}$. After this, each modality obtains a shallow tensor representation, and a simple encoder $f_m = \mathcal{E}_m(X_m)$ is used to extract the feature information of each modality. Here, f_m represents the low-level features of the modality, facilitating subsequent learning in the router network and modality fusion.

Router network. In order to obtain importance weights for various samples across each modality, we design an advanced router network that outputs a refined set of weights conditioned on the input. Formally,

$$f = [f_l, f_v, f_a], \\ W = \mathcal{G}(f; \theta) = \sigma \left(\alpha_i h_i(f) \cdot \frac{1}{t} \right), \quad (1)$$

where $[\cdot, \cdot]$ denotes feature concatenation, σ means scaling function (e.g., softmax) and \mathcal{G} represents the router network, parameterized by θ , composed of multiple linear layers interleaved with batch normalization and activation functions such. Here, $W = \{w_t, w_v, w_a\}$ represents the weights associated with each modality, dynamically learned to capture modality-specific contributions. We introduce a temperature value t in the router network to control the magnitude of the weights and to allow fine-grained adjustments for sharper or softer weight distributions. Of particular note, in contrast to other MoE models that employ single networks as experts, we consider each modality as a

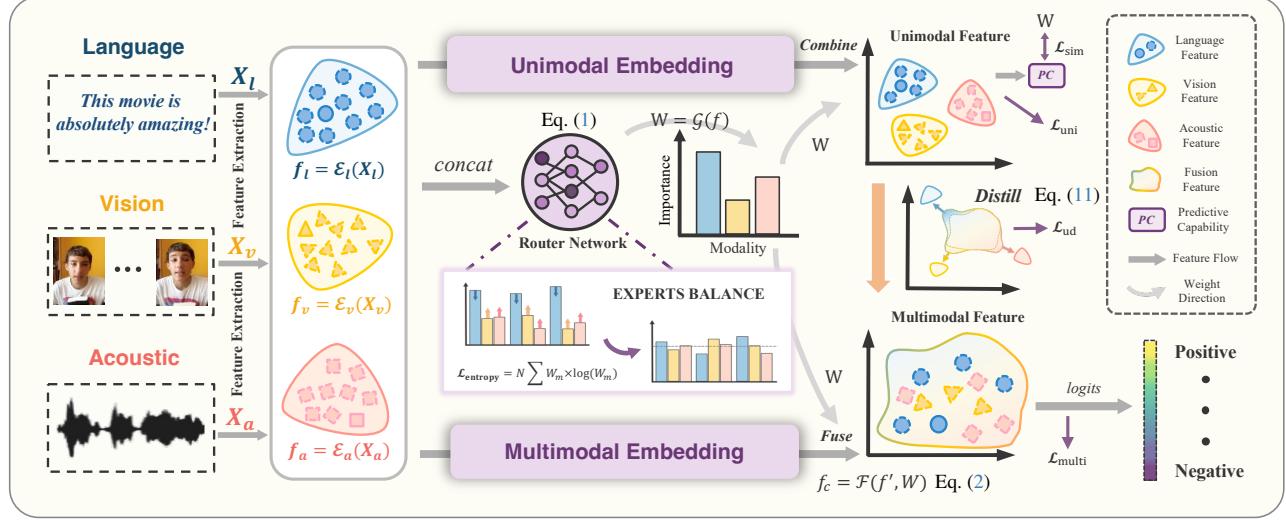


Figure 2. **Architecture illustration** of EMOE. ① After the encoder, Mixture of Modality Experts passes the features through a router network that learns sample-specific weights. These weights guide the fusion of modalities within the multimodal embedding, forming the final multimodal feature (Sec. 3.2). ② The weights also highlight the significance of each modality, allowing features in the unimodal embedding to be prioritized accordingly. Through Unimodal Distillation, critical unimodal feature information is transferred to the multimodal feature, helping retain modality-specific traits (Sec. 3.3).

distinct **expert** and compute weights accordingly, enabling more specialized and powerful capabilities.

Dynamic fusion strategy. After obtaining the corresponding modality weights, fusing the features from different modalities becomes crucial. We design a corresponding transformer on the low-level features to extract their high-level features f'_m , $m \in \{l, v, a\}$, while also addressing the issue of unaligned sequence lengths across different modalities. Next, we focus the modality **fusion methods**, illustrated in Fig. 3. As we introduce the importance weights, weighted summation is the most intuitive form of fusion:

$$\begin{aligned} f_c &= \mathcal{F}(f', W) \\ &= w_l \times f'_l + w_v \times f'_v + w_a \times f'_a, \end{aligned} \quad (2)$$

where f_c means fused feature. Meanwhile, as many existing models rely on concatenation for modality fusion, we additionally design an alternative fusion method based on feature concatenation to ensure compatibility and enhance the generality of our approach:

$$\begin{aligned} f_c &= \mathcal{F}(f', W) \\ &= [w_l \times f'_l, w_v \times f'_v, w_a \times f'_a]. \end{aligned} \quad (3)$$

This paper adopts summation as the default fusion method due to its superior overall performance; further details will be discussed in Sec. 4.2. Based on the fused features obtained in Eq. (2) and Eq. (3), we further use the **fused modality prediction head** to obtain the **prediction result** \hat{y} and compute the loss function using the true result y :

$$\mathcal{L}_{\text{multi}} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (4)$$

Experts balancing mechanism. Although Mixture of Modality Experts learns the basic modal importance capabilities, there still remains some bias in the learned modal weights during specific tests due to the highly individualized nature of the importance distribution across samples. Therefore, we add a **single-modality prediction head**, to calculate the **importance of each modality** for each sample based on the **prediction results**, which can be formulated as:

$$\begin{aligned} c_m &= (y_i - \hat{y}_{i,m})^2, \\ I &= \text{softmax}\left(\frac{1}{c_m + \epsilon}\right), \end{aligned} \quad (5)$$

where \hat{y} represents the **predicted value** for the corresponding single modality, ϵ denotes a very small positive number and I refers to the **importance** coefficients for the three modalities. Since varying modality weights W is designed to best fit the importance of different modalities, we compare it here with the similarity in modality importance:

$$\mathcal{L}_{\text{sim}} = \sum_{m \in \{l, v, a\}} \mathcal{S}(w_m, I_m), \quad (6)$$

where \mathcal{S} means **similarity function**. The computed Weight-Important Similarity assists in learning the classification capabilities of individual modalities, thereby supplementing the modality importance information.

Furthermore, the dynamically weighted fusion approach introduces a new problem. In multimodal learning, there is often a dominant modality that is more effective at conveying information and typically performs better in the early stages of training. However, due to the weighted fusion, the

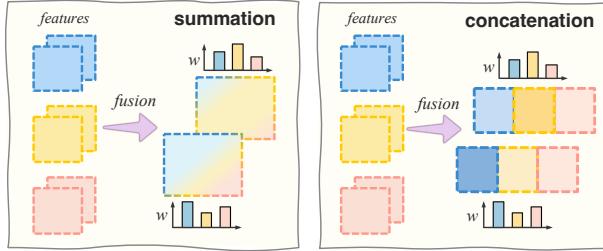


Figure 3. **Fusion Methods Comparison** (Sec. 3.2). *left* depicts the process of summing multimodal features with varying weights; *right* illustrates concatenating features with distinct weights.

model tends to focus excessively on this dominant modality, neglecting the learning of other modalities, which results in an overestimation of its importance. To address this issue, we implement a router entropy loss to discourage the model from over-relying on the dominant modality, thereby preserving its ability to autonomously activate modality-specific experts, as shown below:

$$\mathcal{L}_{\text{entropy}} = N \sum_{m \in \{l, v, a\}} w_m \times \log(w_m + \epsilon), \quad (7)$$

where N represents the number of modality experts, and w_m denotes the weight of the corresponding modality expert. Finally, we combine it with the similarity between modality importance and the weights mentioned to derive the MoME balance loss:

$$\mathcal{L}_{\text{balance}} = \mathcal{L}_{\text{entropy}} + \alpha \mathcal{L}_{\text{sim}}, \quad (8)$$

where α is the scaling factor.

3.3. Unimodal Distillation with Router Selection

Owing to the fact that the aforementioned modality fusion primarily focuses on the homogeneity of the modalities, the modality-specific components may also contain valuable information. Therefore, we leverage the predictive capabilities of single modalities to guide the learning of multimodal features, ensuring that they retain a level of shared commonality while simultaneously capturing the predictive strengths of modality-specific characteristics.

As mentioned in Sec 3.2, we introduce a single-modality prediction head into the model. Since the multimodal approach still relies on the information from each modality, we compute the loss function for single modalities to ensure that they also have predictive capabilities:

$$\begin{aligned} \mathcal{L}_m &= \frac{1}{n} \sum_{i=1}^n |y_i - \tilde{y}_{i,m}|, \quad m \in \{l, v, a\}, \\ \mathcal{L}_{\text{uni}} &= (\mathcal{L}_l + \mathcal{L}_v + \mathcal{L}_a) / 3, \end{aligned} \quad (9)$$

where \mathcal{L}_m means the prediction loss for each modality, \tilde{y} and y represent the unimodal prediction and the truth value.

To learn the specific information in each single modality, we propose a unidirectional knowledge distillation [15] from the multimodal to the unimodal. However, this approach may introduce noise from less significant modalities in certain samples, potentially misguiding the multimodal feature learning. To mitigate this issue, we leverage the modality importance weights W learned by the router network mentioned in Sec 3.2 to weight different unimodal information, emphasizing the modality information that is more relevant to the current sample.

$$z_{\text{uni}} = \mathcal{C}(z_l, z_v, z_a, W) \quad (10)$$

where z_{uni} serves as a combining logit, z_m represents unimodal logits, and \mathcal{C} refers to combine function. The specific form of combination (either summation or concatenation) is determined by the preceding multimodal fusion method. In this way, the strengthened modality effectively takes a leading role in the knowledge distillation:

$$\mathcal{L}_{\text{ud}} = \mathcal{D}(z_{\text{multi}} \rightarrow z_{\text{uni}}), \quad (11)$$

z_{multi} means multimodal logits, \mathcal{D} represents one-way distillation, \rightarrow is distillation direction. Thus, the multimodal prediction results effectively incorporate modality-specific information, drawing on the abilities of unimodal.

3.4. Objective optimization

We integrate the above losses to reach the full objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{balance}} + \lambda_2 \mathcal{L}_{\text{ud}}, \quad (12)$$

where $\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{multi}} + \mathcal{L}_{\text{uni}}$ means the primary task loss, λ_1 control the magnitude of the router network, and λ_2 regulate the importance of modality-specific components.

4. Experiments

Datasets. Considering the generality of our approach for both MER and MIR tasks, we evaluate EMOE on the following three datasets: CMU-MOSI [62], CMU-MOSEI [65] and MIntRec [66]. **CMU-MOSI** consists of 2,199 monologue video samples, with 1,284 for training, 229 for validation, and 686 for testing. Acoustic and visual features are sampled at 12.5 Hz and 15 Hz, respectively. **CMU-MOSEI** includes 22,856 YouTube movie review clips, with 16,326 for training, 1,871 for validation, and 4,659 for testing. Acoustic and visual features are sampled at 20 Hz and 15 Hz. Both datasets have sentiment labels ranging from -3 (*highly negative*) to 3 (*highly positive*). **MIntRec** contains 2,224 samples, including 1,334 for training, 445 for validation, and 445 for testing. It comprises 11 intent types related to emotions or attitudes and 9 related to goal achievement.

Evaluation Metric. For CMU-MOSI and CMU-MOSEI, we follow the previous works [14, 29] to evaluate EMOE

Methods	Setting	ACC ₇ ↑	ACC ₂ ↑	F1↑	MAE↓
EF-LSTM [52]		33.7	75.3	75.2	1.386
LF-DNN [52]		31.5	78.4	78.3	0.972
TFN [63]		31.9	78.8	78.9	0.953
LMF [32]		36.9	78.7	78.7	0.931
MFN [64]		35.6	78.4	78.4	0.964
Graph-MFN [65]		31.5	78.1	78.1	0.970
MCTN [39]	Aligned	33.1	79.3	80.0	0.963
MuLT [44]		35.1	80.0	80.1	0.936
MISA [14] *		41.8	84.2	84.2	0.754
Self-MM [60] *		45.3	84.9	84.9	0.738
MMIM [13] *		45.8	84.6	84.5	0.717
DMD [29] *		46.2	83.2	83.2	0.721
EMOE (Ours)*		47.7	85.4	85.4	0.710
EF-LSTM [52]		31.0	73.6	74.5	1.420
LF-DNN [52]		32.5	78.2	78.3	0.987
TFN [63]		35.3	76.5	76.6	0.995
LMF [32]		31.1	79.1	79.1	0.963
MFN [64]		34.7	80.0	80.1	0.971
Graph-MFN [65]		34.4	79.4	79.2	0.930
MCTN [39]	Unaligned	31.9	77.1	77.3	1.033
MuLT [44]		33.2	80.3	80.3	0.933
MISA [14] *		43.6	83.8	83.9	0.742
Self-MM [60] *		45.7	83.4	83.6	0.724
MMIM [13] *		45.9	83.4	83.4	0.777
DMD [29] *		46.7	84.0	84.0	0.721
EMOE (Ours)*		47.8	85.4	85.3	0.697

Table 1. Comparison on CMU-MOSI dataset. **Bold** is the best. ACC₇, ACC₂ and F1 values are shown as percentages. * indicates BERT-based language features. See details in Sec. 4.1.

by using the metrics: 7-class Accuracy (ACC-7), Binary Accuracy (ACC-2), F1-score (F1), and Mean-absolute Error (MAE). For MIntRec, we follow the standard protocol from the previous work [43, 66], evaluating the results via the following metrics: Accuracy (ACC), F1-score (F1), Precision (P), and Recall (R) for the intent recognition.

Implementation details. For CMU-MOSI and CMU-MOSEI, we utilize 300-dimensional GloVe language features [37] and 768-dimensional BERT-base-uncased hidden states [25]. Facet [2] provides 35 facial action unit visual features, and COVAREP [7] offers 74-dimensional acoustic features. On MIntRec, dimensions for text, visual, and acoustic features are 768, 256, and 768, respectively. The experimental results are obtained by selecting the peak values under the same conditions. Optimal values for α , λ_1 , and λ_2 are set to 0.1, with a fixed temperature of 0.1 based on validation performance. Experiments are conducted on a PyTorch framework using an RTX 4090 GPU with 24GB memory, with a batch size of 16 and training for 50 epochs.

4.1. Comparison with the state-of-the-art

Results on the MER dataset. We compare EMOE with the current state-of-the-art MER methods under the same dataset settings (unaligned or aligned). Tab. 1 and Tab. 2 illustrate the comparison on CMU-MOSI and CMU-MOSEI datasets, respectively. Obviously, our proposed EMOE

Methods	Setting	ACC ₇ ↑	ACC ₂ ↑	F1↑	MAE↓
EF-LSTM [52]		47.4	78.2	77.9	0.620
LF-DNN [52]		51.7	83.5	83.1	0.568
TFN [63]		50.9	80.4	80.7	0.574
LMF [32]		52.3	84.7	84.5	0.564
MFN [64]		50.8	84.0	84.0	0.574
Graph-MFN [65]		51.6	84.6	84.5	0.553
MFM [45]	Aligned	49.4	83.5	83.4	0.590
MuLT [44]		52.3	82.7	82.8	0.572
MISA [14] *		52.3	85.3	85.1	0.543
Self-MM [60] *		53.2	84.5	84.3	0.540
MMIM [13] *		50.1	83.6	83.5	0.580
DMD [29] *		52.4	84.8	84.7	0.546
EMOE (Ours)*		54.1	85.3	85.3	0.536
EF-LSTM [52]		46.3	76.1	75.9	0.594
LF-DNN [52]		52.3	83.7	83.2	0.561
TFN [63]		50.2	84.2	84.0	0.573
LMF [32]		51.9	83.8	83.9	0.565
MFN [64]		51.3	83.2	83.3	0.567
Graph-MFN [65]		51.8	84.2	84.2	0.568
MFM [45]	Unaligned	52.0	82.3	82.5	0.572
MuLT [44]		53.2	84.0	84.0	0.556
MISA [14] *		51.0	84.8	84.8	0.557
Self-MM [60] *		52.9	85.3	84.8	0.535
MMIM [13] *		52.6	81.5	81.3	0.578
DMD [29] *		53.1	84.7	84.7	0.536
EMOE (Ours)*		53.9	85.5	85.5	0.530

Table 2. Comparison on CMU-MOSEI dataset. **Bold** is the best. ACC₇, ACC₂ and F1 values are shown as percentages. * indicates BERT-based language features. See details in Sec. 4.1.

Methods	ACC↑	F1↑	P↑	R↑
MAG-BERT [41]	70.34	68.19	68.31	69.36
MuLT [44]	72.58	69.36	70.73	69.47
MISA [14]	72.36	70.57	71.24	70.41
EMOE (Ours)	72.58	70.73	72.08	70.86

Table 3. Comparison on MIntRec dataset. **Bold** is the best. ACC, F1, P and R values are shown as percentages. Refer to Sec. 4.1.

achieves better accuracy than other MER methods under both unaligned and aligned settings. It is worth highlighting that EMOE exhibits significant improvements in ACC₇, achieving 1.5% increase over DMD [29] (47.7% vs. 46.2%) in the aligned setting and a 1.1% increase (47.8% vs. 46.7%) in the unaligned setting on the CMU-MOSI dataset. Likewise, other metrics also demonstrate varying degrees of improvement. For example, the F1 score increased by 0.5% (85.4% vs. 84.9%) compared to Self-MM [60]. Such improvements are also observed on the CMU-MOSEI dataset. Compare with these state-of-the-art methods that adopt a unified fusion paradigm (such as tensor fusion, graph convolutional networks, and gating mechanisms), the proposed EMOE efficiently learns the importance of different modalities for each sample and performs dynamic modality fusion accordingly. Further, we also retain the predictive capability of unimodal models by leveraging unimodal guidance for multimodal feature learning, which ensures that multimodal learning accounts for both the homogeneity and heterogeneity of the modalities.

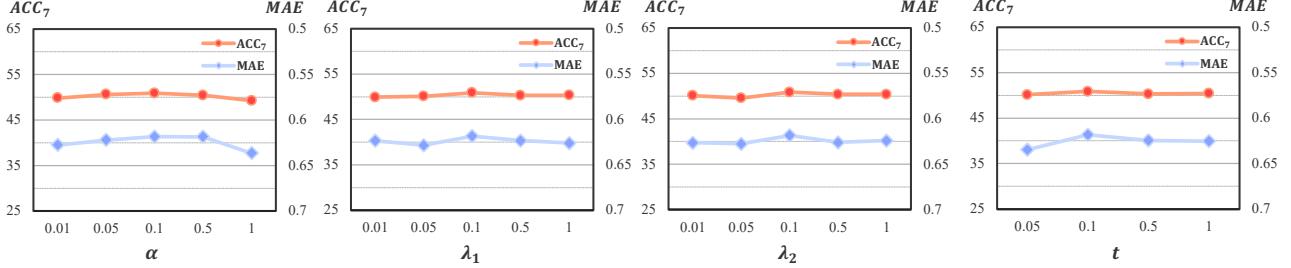


Figure 4. Sensitivity analysis on MER datasets. The results are obtained by varying the value of the corresponding hyperparameter, while fixing the other hyper-parameters to the values adopted in the experiments. Refer to Sec. 4.2.

MoME	UD	ACC_7	ACC_2	F1	MAE
✓		43.0	84.3	84.4	0.751
	✓	46.2	85.2	85.1	0.716
	✓	45.0	84.8	84.7	0.748
✓	✓	47.7	85.4	85.4	0.710

Table 4. Ablation study of the key components in EMOE on MOSI dataset. Please see Sec. 4.2 for details.

Methods	ACC_7	ACC_2	F1
Ours (w/o RE)	45.3	85.1	85.1
Ours (w/o WIS)	46.8	85.3	85.3
Ours	47.8	85.4	85.4

Table 5. Ablation study of the balancing mechanism in EMOE on MOSI dataset. RE is router entropy loss, and WIS is weight-importance similarity loss. Please see Sec. 4.2 for details.

Methods	Setting	ACC_7	ACC_2	F1
Sum	Aligned	47.7	85.4	85.4
		45.9	85.5	85.5
Concat	Unaligned	47.8	85.4	85.3
		47.4	85.4	85.4

Table 6. Fusion methods comparison on MOSI dataset. Sum is summation and concat is concatenation. See Sec. 4.2 for details.

Results on the MIR dataset. To further validate the performance of EMOE, we evaluate the effectiveness on the MIntRec dataset and report the results in Tab. 3. Compare with MISA [14], EMOE demonstrates superior performance, particularly with Precision improving from 71.24% to 72.08%. Other metrics also show varying degrees of improvement. Actually, similar to MER, MIR involves the classification of human intentions. The complexity of the human mental world results in varying sample-level importance across different modalities (language, visual, and acoustic) when expressing intentions. Therefore, the proposed EMOE also achieves outstanding performance when applied to the MIR task, outperforming other methods.

4.2. Ablation study

Quantitative analysis. We evaluate the effects of EMOE’s key components on MOSI dataset, including Mixture of Modality Expert (MoME), Unimodal Distillation (UD). The results are shown in Tab. 4. Our observations are as follows.

Firstly, MoME significantly improves the performance

of MER, indicating that dynamically fusing features from different modalities based on their importance at the sample level enables the model to better utilize useful information from each modality for multimodal information fusion. We also conduct ablation experiments on the expert balance mechanism in MoME. As illustrated in Tab. 5, the weight and contribution similarity loss, along with the routing entropy loss, significantly improve the model’s performance. **Secondly**, we observe that incorporating UD into the model yields notable improvements, particularly in the ACC_7 (47.7% vs. 46.2%) for coarse-grained classification. This suggests that during modal fusion, there is often a focus on the homogeneity between modalities, which can lead to the neglect of valuable coarse-grained classification information provided by individual modalities. UD effectively supplements this modality-specific information, filling in the gaps left by fusion alone. **Thirdly**, to further demonstrate the generality of MoME, we conduct comparative experiments with different fusion strategies (summation and concatenation), as shown in Tab. 6. The results of different fusion strategies are fairly close, suggesting that weighted concatenation can still capture the importance of different modalities, achieving dynamic fusion. Therefore, our approach can serve, to some extent, as a paradigm for solving such problems, with broad applicability.

Sensitivity analysis. To demonstrate the EMOE’s robustness, we conduct the sensitivity analysis for hyperparameters. Since the focus is on the model’s overall performance, we test multiple settings of the loss parameters α , λ_1 , and λ_2 , as well as the temperature value t of the routing network, under both aligned and unaligned configurations on the MOSI and MOSEI datasets. In particular, the sensitivity analysis is conducted by varying the value of the corresponding hyper-parameter, while fixing the other hyper-parameters to the values adopted in the experiments. Fig. 4 presents the averaged results of different parameters across various datasets, illustrating that the overall metrics remain relatively stable. Notably, when the temperature value t is too low, it causes a catastrophic disproportion in the initial weights of the router network, so we excluded those extreme outliers from the analysis. Overall, it is evident that the performance of the proposed approach is not sensitive

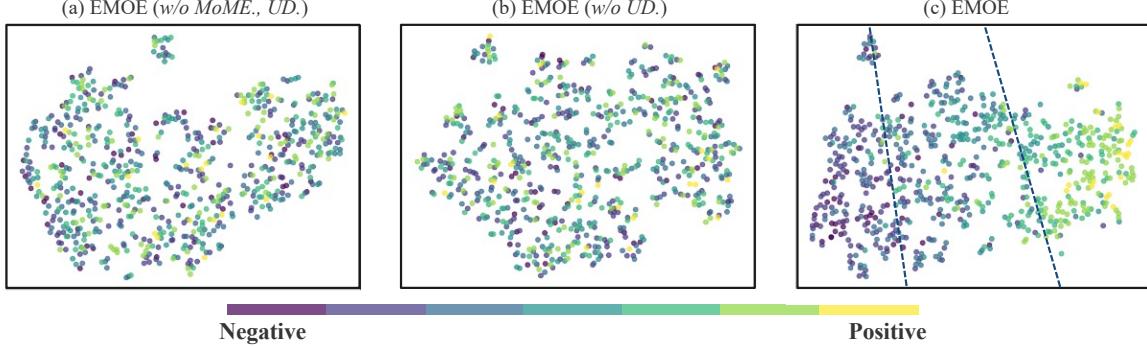


Figure 5. T-SNE visualization of feature distribution on MOSI. The lighter the color, the more positive the emotion. EMOE demonstrates promising performance in the emotion category. Please refer to Sec. 4.2.

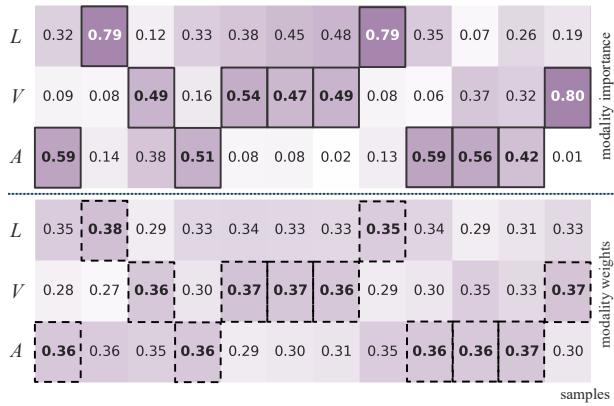


Figure 6. Visualization of modality importance and modality weight on MOSI. The ones with frames represent the maximum values in each column. Please refer to Sec. 4.2.

to the values of the hyper-parameters.

Visualization of feature distribution. To prove the effectiveness of the method, we visualize the feature distributions of EMOE, EMOE (w/o MoME., UD.), and EMOE (w/o UD.) in Fig. 5 for a quantitative comparison. Here, EMOE (w/o MoME., UD.) refers to the baseline without MoME and UD, while EMOE (w/o UD.) represents EMOE with only MoME included. To visualize the feature distribution, we use the test set of the CMU-MOSI dataset as our sample. t-SNE projects selected sample features into 2D space. Under EMOE (w/o MoME., UD.), the distribution is irregular. EMOE (w/o UD.) improves it slightly, while EMOE results show a more compact and consistent distribution aligned with emotional intensity. It is important to note that since this is a regression task, the ideal outcome is a gradient-like distribution, not clustering. Therefore, our results indicate that EMOE enhances feature distinctiveness and better predicts emotions.

Visualization of the modality weight. To demonstrate the effectiveness of MoME, we visualize modality importance and the modality weights across different samples, as shown in Fig. 6. We select some samples from the MOSI dataset and measured each modality’s importance

during fusion based on its single-modality prediction capability. We then compare these contributions to the modality weights obtained through training. Darker color blocks indicate a higher proportion. The results indicate that the predictive efficiency of different weights vary across samples, underscoring the importance of MoME. Additionally, we observe a strong correlation between the learned modality weights and their respective predictive efficiency. This alignment suggests that the model effectively captures and utilizes the relevance of each modality, further demonstrating the efficiency and versatility of our proposed method.

5. Conclusion

In this paper, we explore the issue of inconsistent modality importance across different input samples in the multimodal emotion recognition task. Our work introduces a simple yet effective method for dynamic modality fusion, namely Modality-Specific Enhanced Dynamic Emotion Experts (EMOE). We utilize the Mixture of Modality Experts to calculate the importance weights of each modality, achieving sample-level dynamic fusion. Additionally, we guide multimodal learning through single-modal feature distillation, endowing the prediction results with modality-specific information. The effectiveness of EMOE has been thoroughly validated against numerous popular counterparts across various MER and MIR tasks. We hope this work will serve as a multimodal fusion paradigm, paving the way for future research on related problems.

Acknowledgement. This research is supported by the National Key Research and Development Project of China (2024YFC3308400), the National Natural Science Foundation of China (Grants 62272354, 62361166629, 62176188, 623B2080), the Major Project of Science and Technology Innovation of Hubei Province (2024BCA003) and the Wuhan University Undergraduate Innovation Research Fund Project. The supercomputing system at the Supercomputing Center of Wuhan University supports the numerical calculations in this paper.

References

- [1] Yang Bai, Yucheng Ji, Min Cao, Jinqiao Wang, and Mang Ye. Chat-based person retrieval via dialogue-refined cross-modal alignment. In *CVPR*, 2025. 1
- [2] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *WACV*, pages 1–10. IEEE, 2016. 6
- [3] Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts. *arXiv preprint arXiv:2407.06204*, 2024. 2
- [4] Feiyu Chen, Jie Shao, Shuyuan Zhu, and Heng Tao Shen. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In *CVPR*, pages 10761–10770, 2023. 2
- [5] Ke Chen, Lei Xu, and Huisheng Chi. Improved learning algorithms for mixture of experts in multiclass classification. *Neural networks*, 12(9):1229–1252, 1999. 3
- [6] Tianlong Chen, Xuxi Chen, Xianzhi Du, Abdullah Rashwan, Fan Yang, Huizhong Chen, Zhangyang Wang, and Yeqing Li. Adamv-moe: Adaptive multi-task vision mixture-of-experts. In *ICCV*, pages 17346–17357, 2023. 2, 3
- [7] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep—a collaborative voice analysis repository for speech technologies. In *ICASSP*, pages 960–964. IEEE, 2014. 6
- [8] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *CVPR*, pages 20029–20038, 2023. 3
- [9] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junhong Liu, and Song Guo. Detached and interactive multimodal learning. In *ACM MM*, pages 5470–5478, 2024. 2
- [10] Tommaso Furlanello, Zachary Lipton, Michael Tschanne, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, pages 1607–1616, 2018. 2
- [11] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *NeurIPS*, 33: 3197–3208, 2020. 2
- [12] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *CVPR*, pages 11020–11029, 2020. 2
- [13] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*, 2021. 6
- [14] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *ACM MM*, pages 1122–1131, 2020. 2, 3, 5, 6, 7
- [15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Workshop*, 2015. 2, 5
- [16] Ming Hu, Zeke Xia, Dengke Yan, Zhihao Yue, Jun Xia, Yihao Huang, Yang Liu, and Mingsong Chen. Gitfl: Uncertainty-aware real-time asynchronous federated learning using version control. In *RTSS*, pages 145–157. IEEE, 2023. 2
- [17] Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *CVPR*, 2022.
- [18] Wenke Huang, Mang Ye, Zekun Shi, and Bo Du. Generalizable heterogeneous federated cross-correlation and instance similarity learning. *IEEE PAMI*, 2023. 2
- [19] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *CVPR*, 2023. 1
- [20] Wenke Huang, Mang Ye, Zekun Shi, Guancheng Wan, He Li, Bo Du, and Qiang Yang. A federated learning for generalization, robustness, fairness: A survey and benchmark. *IEEE PAMI*, 2024. 1
- [21] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *ICML*, pages 9226–9259. PMLR, 2022. 3
- [22] Zhijian Huang, Sihao Lin, Guiyu Liu, Mukun Luo, Chaoqiang Ye, Hang Xu, Xiaojun Chang, and Xiaodan Liang. Fuller: Unified multi-modality multi-task 3d perception via multi-level gradient calibration. In *ICCV*, pages 3502–3511, 2023. 2
- [23] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 3
- [24] Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2): 181–214, 1994. 3
- [25] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 6
- [26] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. 3
- [27] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *ICCV*, pages 22214–22224, 2023. 3
- [28] Pingzhi Li, Zhenyu Zhang, Prateek Yadav, Yi-Lin Sung, Yu Cheng, Mohit Bansal, and Tianlong Chen. Merge, then compress: Demystify efficient smoe with hints from its routing policy. *arXiv preprint arXiv:2310.01334*, 2023. 3
- [29] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multi-modal distilling for emotion recognition. In *CVPR*, pages 6631–6640, 2023. 2, 5, 6
- [30] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Ruslan Salakhutdinov. High-modality multi-modal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning. *arXiv preprint arXiv:2203.01311*, 2022. 2

- [31] Zhentao Liu, Min Wu, Weihua Cao, Luefeng Chen, Jianping Xu, Ri Zhang, Mengtian Zhou, and Junwei Mao. A facial expression emotion recognition based human-robot interaction system. *IEEE/CAA Journal of Automatica Sinica*, 4(4): 668–676, 2017. 1
- [32] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *ACL*, 2018. 6
- [33] Prem Melville, Wojciech Gryc, and Richard D Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *KDD*, pages 1275–1284, 2009. 1
- [34] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI*, pages 5191–5198, 2020. 2
- [35] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *ICMI*, pages 169–176, 2011. 2
- [36] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, pages 8238–8247, 2022. 2, 3
- [37] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 6
- [38] Sintija Petrovica, Alla Anohina-Naumeca, and Hazim Kemal Ekenel. Emotion recognition in affective tutoring systems: Collection of ground-truth data. *Procedia Computer Science*, 104:437–444, 2017. 1
- [39] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI*, pages 6892–6899, 2019. 6
- [40] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. pages 108–132. IEEE, 2020. 1
- [41] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pre-trained transformers. In *ACL*, page 2359. NIH Public Access, 2020. 6
- [42] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 3
- [43] Kaili Sun, Zhiwen Xie, Mang Ye, and Huyin Zhang. Contextual augmented global contrast for multimodal intent recognition. In *CVPR*, pages 26963–26973, 2024. 2, 6
- [44] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, 2019. 2, 3, 6
- [45] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *ICLR*, 2019. 6
- [46] An Wang, Xingwu Sun, Ruobing Xie, Shuaipeng Li, Jiaqi Zhu, Zhen Yang, Pinxue Zhao, JN Han, Zhanhui Kang, Di Wang, et al. Hmoe: Heterogeneous mixture of experts for language modeling. *arXiv preprint arXiv:2408.10681*, 2024. 2
- [47] Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-loss-free load balancing strategy for mixture-of-experts. *arXiv preprint arXiv:2408.15664*, 2024. 2
- [48] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, pages 12695–12705, 2020. 2
- [49] Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. *arXiv preprint arXiv:2405.17730*, 2024. 3
- [50] Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *CVPR*, pages 27338–27347, 2024. 3
- [51] Yake Wei, Siwei Li, Ruoxuan Feng, and Di Hu. Diagnosing and re-learning for balanced multimodal learning. In *ECCV*, pages 71–86. Springer, 2025. 2, 3
- [52] Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. Recognizing emotions in video using multimodal dnn feature fusion. In *Challenge-HML*, pages 11–19, 2018. 6
- [53] Xinyu Xie, Yawen Cui, Tao Tan, Xubin Zheng, and Zitong Yu. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *Visual Intelligence*, 2(1): 37, 2024. 2
- [54] Yingjia Xu, Mengxia Wu, Zixin Guo, Min Cao, Mang Ye, and Jorma Laaksonen. Efficient text-to-video retrieval via multi-modal multi-tagger derived pre-screening. *Visual Intelligence*, 3(1):1–13, 2025. 1
- [55] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *ACM MM*, pages 1642–1651, 2022. 2
- [56] Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, et al. Context de-confounded emotion recognition. In *CVPR*, pages 19005–19015, 2023. 2
- [57] Qu Yang, Mang Ye, and Dacheng Tao. Synergy of sight and semantics: visual intention understanding with clip. In *ECCV*, pages 144–160. Springer, 2024. 2
- [58] Qu Yang, Qinghongya Shi, Tongxin Wang, and Mang Ye. Uncertain multimodal intention and emotion understanding in the wild. In *CVPR*, 2025. 2
- [59] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 1
- [60] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *AAAI*, pages 10790–10797, 2021. 6

- [61] Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. *arXiv preprint arXiv:2410.08245*, 2024. 3
- [62] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016. 2, 5
- [63] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, 2017. 2, 6
- [64] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *AAAI*, 2018. 6
- [65] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*, pages 2236–2246, 2018. 2, 5, 6
- [66] Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. Mintrec: A new dataset for multimodal intent recognition. In *ACM MM*, pages 1688–1697, 2022. 2, 5, 6
- [67] Sitao Zhang, Yimu Pan, and James Z Wang. Learning emotion representations from verbal and nonverbal communication. In *CVPR*, pages 18993–19004, 2023. 2
- [68] Xueliang Zhao, Mingyang Wang, Yingchun Tan, and Xianjie Wang. Tgmoe: A text guided mixture-of-experts model for multimodal sentiment analysis. *IJACSA*, 15(8), 2024. 3
- [69] Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, and Xi-angjie Kong. Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, 95:306–325, 2023. 2
- [70] Xingkui Zhu, Yiran Guan, Dingkang Liang, Yuchao Chen, Yuliang Liu, and Xiang Bai. Moe jetpack: From dense checkpoints to adaptive mixture of experts for vision tasks. *arXiv preprint arXiv:2406.04801*, 2024. 3
- [71] Yun Zhu, Nevan Wicher, Chu-Cheng Lin, Xinyi Wang, Tianlong Chen, Lei Shu, Han Lu, Canoee Liu, Liangchen Luo, Jindong Chen, et al. Sira: Sparse mixture of low rank adaptation. *arXiv preprint arXiv:2311.09179*, 2023. 3